

# PRIMES IN THE INTERVALS BETWEEN PRIMES SQUARED

KOLBJØRN TUNSTRØM

**ABSTRACT.** The set of short intervals between consecutive primes squared has the pleasant—but seemingly unexploited—property that each interval  $s_k := \{p_k^2, \dots, p_{k+1}^2 - 1\}$  is fully sieved by the  $k$  first primes. Here we take advantage of this essential characteristic and present evidence for the conjecture that  $\pi_k \sim |s_k|/\log p_{k+1}^2$ , where  $\pi_k$  is the number of primes in  $s_k$ ; or even stricter, that  $y = x^{1/2}$  is both necessary and sufficient for the prime number theorem to be valid in intervals of length  $y$ . In addition, we propose and substantiate that the prime counting function  $\pi(x)$  is best understood as a sum of correlated random variables  $\pi_k$ . Under this assumption, we derive the theoretical variance of  $\pi(p_{k+1}^2) = \sum_{j=1}^k \pi_j$ , from which we are led to conjecture that  $|\pi(x) - \text{li}(x)| = O(\sqrt{\text{li}(x)})$ . Emerging from our investigations is the view that the intervals between consecutive primes squared hold the key to a furthered understanding of the distribution of primes; as evidenced, this perspective also builds strong support in favour of the Riemann hypothesis.

## 1. INTRODUCTION

An important theme in analytic number theory is the distribution of primes in short intervals. Notably, it is still an open question for what exact interval lengths the prime number theorem is valid or breaks down, as it eventually does for short enough intervals. The problem, formally stated, is to identify for which functions  $y = y(x)$ , as  $x \rightarrow \infty$ ,

$$(1) \quad \pi(x+y) - \pi(x) \sim \frac{y}{\log x}.$$

Currently, the gulf between the known upper and lower bounds of  $y$  is huge. The best unconditional estimate of the upper bound, proved by Heath-Brown [Heath-Brown 1988], is  $y = x^{7/12+\epsilon(x)}$ , where  $\epsilon(x) \rightarrow 0$  as  $x \rightarrow \infty$ , an estimate that is lowered to  $y = x^{1/2+\epsilon}$  assuming the Riemann hypothesis (RH). In contrast, Selberg, also assuming RH, proved that as long as  $y/(\log x)^2 \rightarrow \infty$  when  $x \rightarrow \infty$ , then (1) holds for *almost all*  $x$ , understood in the sense that the set of  $x \in [0, X]$  for which (1) is not true is  $o(X)$  as  $X \rightarrow \infty$ . Unconditionally, this almost-all result has been proven to hold with  $y = x^{1/6-\epsilon}$  [Zaccagnini 1998]. It was long thought that Selberg's result would be true even in the case of all  $x$  [Granville 1995b], but Maier [Maier 1985] made it clear that there are indeed infinitely many exceptions to Selberg's result, for any function of the form  $y = (\log x)^\lambda$ , with  $\lambda > 1$ . It is not known whether there exist exceptions that are larger than  $(\log x)^\lambda$ , but it has been conjectured that  $y > x^\epsilon$  will suffice for (1) to hold for all  $x$ , see e.g. [Granville 1995b] or [Soundararajan 2007].

In this paper we draw attention to the short intervals between consecutive primes squared, defined by  $s_k := \{p_k^2, \dots, p_{k+1}^2 - 1\}$  for  $k \geq 1$ . These intervals naturally occur in the context of the sieve of Eratosthenes, and in particular, each  $s_k$  has

the specific quality of being fully sieved by the  $k$  first primes; any element in  $s_k$  is either divisible by some  $p \in \mathcal{P}_k := \{p_1, \dots, p_k\}$  or else is a prime  $p \notin \mathcal{P}_k$ . In addition, the exact distribution of primes in  $s_k$  is in its entirety build up of the periodic sequences

$$\rho_k(n) := \begin{cases} p_k & \text{if } p_k \mid n, \\ 1 & \text{otherwise,} \end{cases}$$

which we visualise for the specific example of  $s_3$  by the following table:

$n$	25	26	27	28	<b>29</b>	30	<b>31</b>	32	33	34	35	36	<b>37</b>	...	48
$\rho_1(n)$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	...	$p_1$
$\rho_2(n)$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	...	$p_2$
$\rho_3(n)$	$p_3$	1	1	1	1	$p_3$	1	1	1	1	$p_3$	1	1	...	1

Together with the fact that these intervals make up a complete subdivision of the natural numbers, these insights allow us to argue for the range in which the prime number theorem is valid, as well as providing a rudimentary explanation of the behaviour of the prime counting function  $\pi(x)$ . Specifically, we present a heuristic as to why the following conjecture should hold:

**Conjecture.** *The choice of  $y = x^{1/2}$  is necessary and sufficient for*

$$\pi(x+y) - \pi(x) \sim \frac{y}{\log(x+y)}$$

*to hold for all  $x$  as  $x \rightarrow \infty$ .*

Furthermore, we present evidence that the set of  $\pi_k$ s satisfies a conjecture of Montgomery and Soundararajan [Montgomery and Soundararajan 2004], that the number of primes in  $s_k$ , denoted  $\pi_k$ , is expected to follow a normal distribution with mean

$$\mu_k = \frac{|s_k|}{\log p_{k+1}^2}$$

and standard deviation

$$\sigma_k = \frac{\sqrt{|s_k| \log(p_{k+1}^2 / |s_k|)}}{\log p_{k+1}^2}.$$

Because of the periodic structure of the sequences  $\rho_k(n)$ , the number of primes in two intervals  $s_i$  and  $s_j$  are not independent, but correlated. Following from this, we demonstrate that:

**Remark.** The prime counting function  $\pi(x)$  behaves —and should be understood— as a sum of *correlated* random variables  $\pi_k$ , each normally distributed with mean  $\mu_k$  and standard deviation  $\sigma_k$ .

Building on this insight, we formulate a random model of the primes, where the random variables are the number of primes in each interval  $s_k$ , and where these variables derive from drawing random translations of the sequences  $\rho_k(n)$ . This construction naturally preserves the observed correlations between  $\pi_k$ s. Moreover, assuming this model, we calculate the theoretical variance of  $\pi(p_{k+1}^2) = \sum_{j=1}^k \pi_j$ , which in turn suggests the conjecture

**Conjecture.** *The error term in the prime number theorem satisfies*

$$|\pi(x) - \text{li}(x)| = O\left(\sqrt{\text{li}(x)}\right).$$

In sum, it emerges that the apparent random behaviour of the primes and the prime counting function  $\pi(x)$  can be fundamentally understood in terms of the intervals  $s_k$  and the underlying periodic sequences  $\rho_k(n)$ . Through Koch's equivalence criterium [von Koch 1901], our results also strongly support the correctness of the Riemann hypothesis, and as such might pave the way towards a complete technical proof.

## 2. NOTATION AND DEFINITIONS

We write the set consisting of the  $k$  first primes as  $\mathcal{P}_k := \{p_1, \dots, p_k\}$ , the intervals between consecutive primes squared as  $s_k := \{p_k^2, \dots, p_{k+1}^2 - 1\}$ , and the length of  $s_k$  as  $l_k := |s_k| = p_{k+1}^2 - p_k^2$ , where  $k \geq 1$ . As usual, the number of primes less than or equal to  $x$  is given by  $\pi(x)$ , but in addition, since the number of primes in  $s_k$  will appear frequently, we establish the shorthand notation

$$\pi_k := \pi(p_{k+1}^2) - \pi(p_k^2).$$

For the same reason, in the case when we apply the logarithmic integral,

$$\text{li}(x) := \int_2^x \frac{dt}{\log t},$$

to the interval  $s_k$ , we write

$$\text{li}_k := \text{li}(p_{k+1}^2) - \text{li}(p_k^2).$$

Moreover, we need notation for the expected number of primes in  $s_k$ —understood as the expected number of coprimes to  $p_k\#$  in a random interval of length  $l_k$  (where  $p_k\# := \prod_{p \in \mathcal{P}_k} p$  is the primorial of  $p_k$ ). The probability of a random integer being coprime to  $p_k\#$  is given by the Euler product  $\prod_{p \in \mathcal{P}_k} \left(1 - \frac{1}{p}\right)$ , and multiplying this by  $l_k$  produces the expected number of primes in  $s_k$ , which we denote

$$(2) \quad \tilde{\pi}_k := l_k \cdot \prod_{p \in \mathcal{P}_k} \left(1 - \frac{1}{p}\right).$$

We now use this definition to construct a probabilistic prime counting function  $\tilde{\pi}(x)$ —the expected number of primes less than or equal to  $x$ . By assuming  $k$  to be the integer such that  $p_k^2 \leq x < p_{k+1}^2$ , we define  $\tilde{\pi}(x)$  simply as the sum taken over the individual estimates  $\tilde{\pi}_j$ ,  $1 \leq j \leq k$ , namely

$$(3) \quad \tilde{\pi}(x) := \sum_{j=1}^{k-1} \tilde{\pi}_j + \frac{x - p_k^2}{l_k} \tilde{\pi}_k.$$

The last term on the right side adjusts for the fact that  $x$  in general reaches only partially into the last interval  $s_k$ .

Finally, we emphasise the fact that the distribution of primes within any interval  $s_k$  can be viewed as a construction of  $k$  periodic sequences. To make this structure

apparent, we first define an arithmetic function that picks out the integers  $n$  coprime to a given prime  $p_k$ ,

$$\rho_k(n) := \begin{cases} p_k & \text{if } p_k \mid n, \\ 1 & \text{otherwise.} \end{cases}$$

Then we apply this definition to construct a second arithmetic function that locates all  $n$  coprime to  $p_k\#$ ,

$$R_k(n) := \prod_{1 \leq i \leq k} \rho_i(n).$$

By these definitions  $R_k(n) = 1$  whenever  $(n, p_k\#) = 1$ , and both  $\rho_k(n)$  and  $R_k(n)$  are periodic, satisfying for any integer  $m$  the equalities

$$\rho_k(n + mp_k) = \rho_k(n) \quad \text{and} \quad R_k(n + mp_k\#) = R_k(n).$$

Since  $s_k$  is sieved completely by the  $k$  first primes, the primes within  $s_k$  align with the 1s in  $R_k$ , and therefore the number of primes in  $s_k$  equals the number of 1s in  $R_k$  across the interval. We visualise this for the specific example of  $s_3$  by the following table:

$n$	25	26	27	28	<b>29</b>	30	<b>31</b>	32	33	34	35	36	<b>37</b>	...	48
$\rho_1(n)$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	...	$p_1$
$\rho_2(n)$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	...	$p_2$
$\rho_3(n)$	$p_3$	1	1	1	1	$p_3$	1	1	1	1	$p_3$	1	1	...	1
$R_3(n)$	$p_3$	$p_1$	$p_2$	$p_1$	<b>1</b>	$p_1 p_2 p_3$	<b>1</b>	$p_1$	$p_2$	$p_1$	$p_3$	$p_1 p_2$	<b>1</b>	...	$p_1 p_2$

But in general, if we pick a random interval  $\tilde{s}_k$  of length  $l_k$ , the 1s in  $R_k$  across  $\tilde{s}_k$  are not necessarily primes; all we know is that they are coprime to  $p_k\#$ . And because  $R_k$  is periodic with period  $p_k\#$ , the number of coprimes to  $p_k\#$  in  $\tilde{s}_k$  can take any out of  $p_k\#$  (non-unique) values; the probabilistic counting function  $\tilde{\pi}_k$  is the mean value of these.

The last statement can be made explicit by applying sieve notation. Consider therefore an arbitrary interval  $A$  and denote the number of coprimes to  $p_k\#$  in  $A$  by

$$S(A, p_k\#) := |\{n : n \in A, R_k(n) = 1\}|.$$

It follows from the periodicity of  $R_k$  that  $S(A, p_k\#)$  is periodic as well; if  $A^j$  denotes  $A$  left-shifted  $j$  times, the equality

$$S(A^{m \cdot p_k\#}, p_k\#) = S(A, p_k\#)$$

holds for any integer  $m$ . Using this notation, it is obvious that we can restate the probabilistic prime counting function  $\tilde{\pi}_k$  as the mean value taken over the sample space

$$\Omega_k := \{S(s_k^j, p_k\#)\}_{j=0}^{p_k\#-1}.$$

In other words,

$$\tilde{\pi}_k = \frac{1}{p_k\#} \sum_{j=0}^{p_k\#-1} S(s_k^j, p_k\#).$$

The actual number of primes in  $s_k$  coincides with the specific element in  $\Omega_k$  corresponding to  $j = 0$ ,

$$\pi_k = S(s_k^0, p_k\#) = S(s_k, p_k\#).$$

The prime counting function  $\pi_k$  is therefore just one instance out of  $p_k\#$  elements in the sample space underlying  $\tilde{\pi}_k$ . The crucial point, nonetheless, is that all elements in  $\Omega_k$ ,  $\pi_k$  inclusive, stem from the same underlying structure of  $k$  periodic sequences.

### 3. APPROXIMATE EXPRESSIONS FOR $\tilde{\pi}_k$ AND $\tilde{\pi}(x)$

From (2), we know that the expected number of primes in the interval  $s_k$  can be expressed exactly in terms of

$$\tilde{\pi}_k = l_k \cdot \prod_{p \in \mathcal{P}_k} \left(1 - \frac{1}{p}\right).$$

Applying Merten's product theorem, we attain the following approximation:

**Lemma 3.1.**

$$(4) \quad \tilde{\pi}_k \sim 2e^{-\gamma} \frac{l_k}{\log p_{k+1}^2}.$$

*Proof.* Assume that  $x$  is a real number within the interval  $s_k$ , so that  $p_k^2 \leq x < p_{k+1}^2$ . Then we can state Merten's product theorem [Mertens 1874] as

$$(5) \quad \prod_{p \in \mathcal{P}_k} \left(1 - \frac{1}{p}\right) = e^{-\gamma+\delta} \frac{1}{\log \sqrt{x}} = 2e^{-\gamma+\delta} \frac{1}{\log x},$$

where  $\gamma$  is the Euler–Mascheroni constant and  $\delta$  is a measure of the uncertainty of the approximation, satisfying

$$(6) \quad |\delta| < \frac{4}{\log(\sqrt{x}+1)} + \frac{2}{\sqrt{x} \log \sqrt{x}} + \frac{1}{2\sqrt{x}}.$$

By combining (2) and (5), it follows immediately that

$$\tilde{\pi}_k = 2e^{-\gamma+\delta} \frac{l_k}{\log x}.$$

Additionally, (6) implies that in order to minimize the error we should choose  $x$  as large as possible—that is,  $x = p_{k+1}^2 - \epsilon$ , where  $\epsilon > 0$  is infinitesimal—by which we obtain (4).  $\square$

Note that by the choice of  $x = p_{k+1}^2$  in the proof above, the estimate of  $\tilde{\pi}_k$  reflects the characteristic property of the sieving process; sieving by the  $k$  first primes removes all composites less than  $p_{k+1}^2$ .

Similarly, we obtain an approximate expression for  $\tilde{\pi}(x)$ :

**Lemma 3.2.**

$$\tilde{\pi}(x) \sim 2e^{-\gamma} \text{li}(x).$$

*Proof.* First, applying Lemma 3.1 to (3) we immediately have that

$$\tilde{\pi}(x) \sim 2e^{-\gamma} \left( \sum_{j=1}^{k-1} \frac{l_j}{\log p_{j+1}^2} + \frac{x - p_k^2}{l_k} \frac{l_k}{\log p_{k+1}^2} \right).$$

The expression within parentheses is nothing but a Riemann sum, which in the continuum limit can be approximated by the logarithmic integral  $\text{li}(x)$ : The logarithmic integral taken over the interval  $s_k$  satisfies

$$\frac{l_k}{\log p_k^2} > \text{li}_k > \frac{l_k}{\log p_{k+1}^2},$$

and also,

$$\frac{l_k}{\log p_k^2} \sim \frac{l_k}{\log p_{k+1}^2}.$$

We therefore have that

$$\tilde{\pi}_k \sim 2e^{-\gamma} \text{li}_k,$$

and eventually,

$$\tilde{\pi}(x) \sim 2e^{-\gamma} \left( \sum_{j=1}^{k-1} \text{li}_j + \frac{x - p_k^2}{l_k} \text{li}_k \right) = 2e^{-\gamma} \text{li}(x).$$

□

We emphasise here again that the above results are derived under the sole assumption that  $s_k$  is sieved by the  $k$  first primes. In this case, the best we can do is to assign a uniform probability of finding a prime in any position across  $s_k$ . This probability is given exactly by  $\prod_{p \in \mathcal{P}_k} \left(1 - \frac{1}{p}\right)$ , or approximately by  $2e^{-\gamma} / \log p_{k+1}^2$ . Obviously, there is more to say about  $s_k$ —illustrated with a few examples in the next section—and this additional information is what eventually will close the gap between  $\tilde{\pi}_k$  and  $\pi_k$ , where the latter is anticipated to satisfy

$$\pi_k \sim \frac{l_k}{\log p_{k+1}^2}.$$

It is also worth noting that the expression  $2e^{-\gamma} \text{li}(x)$  emerges as the continuum approximation to the discrete sum  $2e^{-\gamma} \sum_{j=1}^k l_j / \log p_{j+1}^2$ , and not the other way around. This fact hints to the possibility that even the prime counting function  $\pi(x)$  is best approximated by  $\sum_{j=1}^k l_j / \log p_{j+1}^2$  rather than  $\text{li}(x)$ . We examine this in closer detail in Section 6.

#### 4. SHRINKING THE GAP BETWEEN $\tilde{\pi}_k$ AND $\pi_k$

Recall from Section 2 that the probabilistic prime counting function  $\tilde{\pi}_k$  can be stated as

$$\tilde{\pi}_k = \frac{1}{p_k \#} \sum_{j=0}^{p_k \# - 1} S(s_k^j, p_k \#),$$

while  $\pi_k$  is given by

$$\pi_k = S(s_k^0, p_k \#).$$

To close in on an estimate for  $\pi_k$ , we need to shrink the sample space  $\Omega_k$  in such a way that it still contains  $S(s_k^0, p_k \#)$ . Of course,  $S(s_k^0, p_k \#)$  is completely determined by where in the natural numbers the interval  $s_k$  is situated (allowing for shifts that are multiples of  $p_k \#$ ), so the constraints on  $\Omega_k$  must reflect this fact. While defining the right constraints is an essential part of sieve theory, our point

here is only to illustrate with a few numerical examples how imposing constraints affect the probabilistic estimate  $\tilde{\pi}_k$ .

The most obvious constraint is that all elements in  $R_k(n)$  must be strictly smaller than  $p_{k+1}^2$  whenever  $n \in s_k$ . We can observe the effect of this constraint by expanding the Euler product (2) in terms of the the Möbius function, which is defined by

$$(7) \quad \mu(n) := \begin{cases} 1 & \text{if } n = 1, \\ (-1)^k & \text{if } n \text{ is a product of } k \text{ distinct primes,} \\ 0 & \text{if } n \text{ has one or more repeated prime factors.} \end{cases}$$

It follows that we can rewrite (2) as

$$\tilde{\pi}_k = l_k \cdot \sum_{d|p_k\#} \frac{\mu(d)}{d}.$$

Adding the constraint produces a truncated version, which we denote

$$\tau_k = l_k \cdot \sum_{\substack{d|p_k\# \\ d < p_{k+1}^2}} \frac{\mu(d)}{d}$$

Numerically—as seen in Figure 1—we verify that the truncated expression amounts to a significant reduction of the sample space  $\Omega_k$ , resulting in  $\tau_k/(l_k/\log p_{k+1}^2) \approx 1.03$ , as opposed to  $\tilde{\pi}_k/(l_k/\log p_{k+1}^2) \sim 2e^{-\gamma} \approx 1.12$ . In addition, the error term, which is  $O(2^k)$  in the case of  $\tilde{\pi}_k$ , reduces to  $O(k^{2.32})$  in the case of  $\tau_k$ .

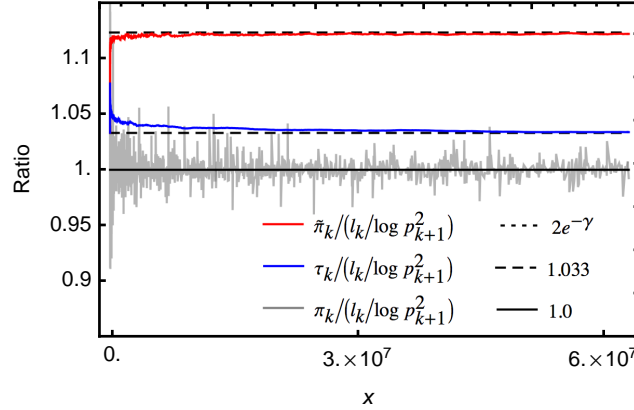


FIGURE 1. The ratios of  $\pi_k$ ,  $\tilde{\pi}_k$ , and  $\tau_k$  to  $l_k/\log p_{k+1}^2$ . The values are plotted at  $x = p_{k+1}^2$ ,  $1 \leq k \leq 1000$ . The three horizontal lines illustrate the limits tended to by each ratio.

Our second example is a set of constraints on the internal structure of  $s_k$ : Within  $s_k$ , the first appearance of a composite divisible by  $p_i$ , for  $i < k$ , is in one of the positions  $p_i - m + 1$ , where  $m > 0$  is an element in the residue class of  $n^2 \pmod{p_i}$ ,  $j > i$  ( $m = 0$  is not included, since the first position in  $s_k$  is always occupied by

$p_k^2$ ). For instance, given the primes  $p_1, \dots, p_6$ , these are the possible positions of first appearance when  $k > 6$ :

Prime	Candidate positions of first appearance in $s_k$
$p_1$	2
$p_2$	3
$p_3$	2, 5
$p_4$	4, 6, 7
$p_5$	3, 7, 8, 9, 11
$p_6$	2, 4, 5, 10, 11, 13

The outcome is that these constraints bound the possible positions of sequences  $\rho_i(n)$  across  $s_k$ ,  $1 \leq i \leq k$ , effectively halving the sample space  $\Omega_k$ .

To get an impression of how these constraints impact the probabilist prime counting function  $\tilde{\pi}_k$ , we randomly sample sequences of  $\rho_i$  of length  $l_k$ ,  $1 \leq i \leq k$ , with first appearance of  $p_i$  in each sequence constrained as explained above, and count the number of 1s in the resulting  $R_k$ —denoted  $\eta_k$ . We observe in Figure 2 that the mean value of  $\eta_k$  in fact appears to lie slightly above  $\tilde{\pi}_k$ , while still satisfying  $\eta_k \sim \tilde{\pi}_k$ . Note that in the figure we plot  $\sum_{j=1}^k \eta_j$  rather than  $\eta_k$  as the accumulation of biases is more visible than the individual biases in each interval. It appears therefore that the statistical properties of the reduced sample space do not change significantly from the original one, which can be understood from the fact that these constraints do not restrain the position of  $s_k$  within the natural numbers.

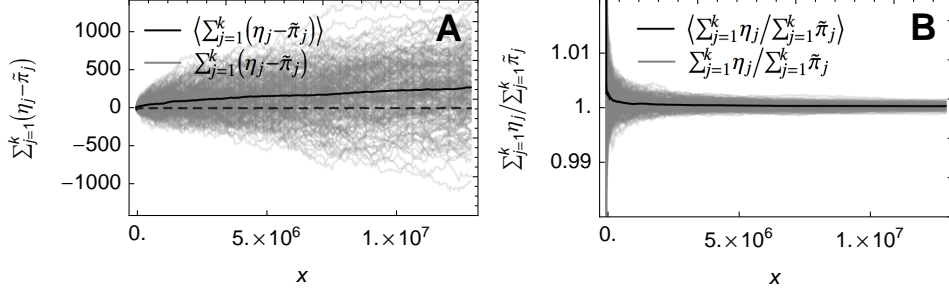


FIGURE 2. Comparison of  $\sum_{j=1}^k \eta_j$  and  $\sum_{j=1}^k \tilde{\pi}_j$ , with values plotted at  $x = p_{k+1}^2$ ,  $1 \leq k \leq 500$ . A) 200 samples of  $\sum_{j=1}^k (\eta_j - \tilde{\pi}_j)$  (grey) shown together with the mean value  $\langle \sum_{j=1}^k (\eta_j - \tilde{\pi}_j) \rangle$  (black), demonstrating a statistical bias of  $\eta_k > \tilde{\pi}_k$ . B) 200 samples of  $\sum_{j=1}^k \eta_j / \sum_{j=1}^k \tilde{\pi}_j$  (grey) shown together with the mean value  $\langle \sum_{j=1}^k \eta_j / \sum_{j=1}^k \tilde{\pi}_j \rangle$  (black), indicating that  $\eta_k \sim \tilde{\pi}_k$ .

## 5. CONJECTURES ON THE DISTRIBUTION OF PRIMES IN SHORT INTERVALS

The prime number theorem, proven independently by Hadamard and de la Vallée-Poussin in 1896, is usually stated by either of the equivalent relations

$$\pi(x) \sim \frac{x}{\log x} \quad \text{or} \quad \pi(x) \sim \text{li}(x).$$



The results so far, however, motivates an alternative formulation of the prime number theorem, in the form of the conjecture:

**Conjecture 1.** *As  $k \rightarrow \infty$ , we have that*

$$\pi_k \sim \frac{l_k}{\log p_{k+1}^2}.$$

From this perspective—as in the probabilistic case in Section 3—we should view  $\pi(x) \sim \text{li}(x)$  as originating from a continuum approximation of  $\sum_{j=1}^k l_j / \log p_{j+1}^2$ .

Proving Conjecture 1 rigorously is out of reach for this paper, but we present a heuristic justification that hopefully can seed a complete proof. Specifically, we argue that the following conjecture holds, which in turn implies Conjecture 1:

**Conjecture 2.** *The choice of  $y = x^{1/2}$  is necessary and sufficient for*

$$(8) \quad \pi(x+y) - \pi(x) \sim \frac{y}{\log(x+y)}$$

*to hold for all  $x$  as  $x \rightarrow \infty$ .*

To start with, let us consider what happens to an interval that grows slower than  $y = x^{1/2}$ . For this purpose we define  $y_\epsilon = y_\epsilon(x)$  to be a function that grows arbitrarily slower than  $y$ . Naturally, no matter how large  $y_\epsilon$  is initially, as  $x$  increases,  $y_\epsilon$  eventually becomes infinitesimal compared to  $y$ . This has the effect that for any given integer  $q > 0$ , we can find an  $x$  such that  $y_\epsilon \leq y/2^q$ . It then follows from Bertrand's postulate—which says there is always a prime  $p$  such that  $n < p < 2n$ —that

$$\pi(y) - \pi(y_\epsilon) \geq \pi(y) - \pi(y/2^q) = q.$$

And because  $q$  can be arbitrarily large,

$$\lim_{x \rightarrow \infty} \pi(y) - \pi(y_\epsilon) = \infty.$$

Therefore, across  $s_k$  we will always find that  $p_k \leq y < p_{k+1}$ , while  $y_\epsilon$  lags behind and ultimately turns infinitesimal relative to infinitely many primes smaller than  $p_k$ .

Let us explore the consequence of the last statement: As described earlier, an essential property of the distribution of primes in  $s_k$  is that we can view it as a construction of the periodic sequences  $\rho_j(n)$ ,  $1 \leq j \leq k$ . From this perspective, it is clear that if we want to sample correctly the distribution of primes within  $s_k$ , we need to choose  $y$  such that the interval  $[x, x+y]$  spans at least the largest underlying period, that is,  $y \geq p_k$  whenever  $x \in s_k$ ; the slowest growing function that satisfies this criteria is  $y = x^{1/2}$ . The interval  $[x, x+y_\epsilon]$ , on the other hand, is in general only long enough to provide valid samples from the distribution of coprimes to  $p_m\#$ , where  $p_m$  is the greatest prime smaller or equal to  $y_\epsilon$  across  $s_k$ . Since, from our argument above,  $p_m$  eventually grows infinitesimal relative to arbitrarily many primes smaller than  $p_k$ , we should expect  $[x, x+y_\epsilon]$  to ultimately move repeatedly across regions where the density of primes deviates significantly from that predicted by the prime number theorem.

What this suggests is that  $y = x^{1/2}$  is the sharp barrier below which the prime number theorem breaks down for all  $x$ , contradicting previous conjectures that this barrier lies close to  $y = x^\epsilon$ ,  $\epsilon > 0$  [Granville 1995b, Soundararajan 2007]. Furthermore, taking this reasoning to its conclusion, it seems logical to conjecture

a variation of Maier's theorem [Maier 1985] applied to any interval growing slower than  $x^{1/2}$ :

**Conjecture 3.** *Let  $y_\epsilon = y_\epsilon(x)$  satisfy  $\lim_{x \rightarrow \infty} y_\epsilon/y = 0$ , where  $y = x^{1/2}$ . Then*

$$\limsup_{x \rightarrow \infty} \frac{\pi(x + y_\epsilon) - \pi(x)}{y_\epsilon / \log x} > 1 \quad \text{and} \quad \liminf_{x \rightarrow \infty} \frac{\pi(x + y_\epsilon) - \pi(x)}{y_\epsilon / \log x} < 1.$$

So far, we have argued that  $y = x^{1/2}$  is a necessary condition for (8) to hold for all  $x$ . To substantiate that  $y = x^{1/2}$  is also sufficient, we make the observation that the lengths of the intervals  $s_k$  are of the form  $l_k = 2p_{k+1}g_k - g_k^2$ , where  $g_k := p_{k+1} - p_k$ , and hence lie on the curves  $2\sqrt{x}g - g^2$ , with  $g = 2n$ ,  $n \geq 1$ . Therefore,  $l_k$  grows as  $O(x^{1/2})$ . Let us further define  $y_\delta = y_\delta(x)$  to be a function that grows arbitrarily faster than  $y = x^{1/2}$ , so that  $\lim_{x \rightarrow \infty} y_\delta/y = \infty$ . As a consequence, the interval  $[x, x + y_\delta]$  will eventually cover arbitrarily many intervals  $s_k$ ; that is, for any  $m$ , we can always choose  $k$  so that  $y_\delta(p_k^2) \geq \sum_{i=k}^{k+m} l_i$ . The problem we encounter now, however, is that the estimate

$$\pi(p_k^2 + y_\delta) - \pi(p_k^2) \sim \frac{y_\delta}{\log(p_k^2 + y_\delta)}$$

assumes a uniform distribution of primes in the interval  $[p_k^2, p_k^2 + y_\delta]$ , given by the density of primes in  $s_m$ . But this assumption is inaccurate; the largest intervals across which we can suppose a uniform distribution are the intervals  $s_k$ . Subdividing the natural numbers into intervals of length  $y_\delta$  and estimating the number of primes up to  $x$  from these results in an underestimate of the prime counting function  $\pi(x)$ . At its most extreme, this is exemplified by the estimate  $\pi(x) \sim x/\log x$ , which is well known to be an inferior guess of the number of primes up to  $x$  compared to  $\pi(x) \sim \text{li}(x)$ , a fact that was established in 1899 by de la Vallée-Poussin [de la Vallée-Poussin 1899].

One final note about this heuristic. We have only argued for what lengths  $y$  can take, in order for the interval  $[x, x+y]$  to correctly sample the distribution of primes, but not what is the correct limiting value of  $\pi(x+y) - \pi(x)$ . The asymptotic limit in Conjecture 2 should therefore be considered as an assumption of the heuristic as well.

## 6. NUMERICAL RESULTS ON THE DISTRIBUTION OF PRIMES

With the purpose of providing clues about how to understand the apparent random behaviour of the primes, we here examine the distribution of primes both locally and globally. We start by noting that since  $l_k = 2p_{k+1}g_k - g_k^2$ , where  $g_k = p_{k+1} - p_k$ , we have that

$$\pi_k \sim \frac{l_k}{\log p_{k+1}^2} = \frac{2p_{k+1}g_k - g_k^2}{\log p_{k+1}^2}.$$

Naturally,  $g_k$  is not unique, suggesting that we can subdivide the set of intervals  $s_k$  according to what the corresponding gap  $g_k$  is. Hence, the subset of intervals  $s_k$  all with corresponding gap  $g$  will be distributed along the curve

$$\frac{2\sqrt{x}g - g^2}{\log x},$$

where  $g$  takes the values  $2n$ ,  $n \geq 1$ . This structure becomes apparent when plotting the values of  $\pi_k$  as a function of  $p_{k+1}^2$ , as illustrated in Figure 3, where all values of  $\pi_k$  are plotted for  $1 \leq k \leq 6 \times 10^5$ .

While this perspective clearly reveals how the intervals  $s_k$  define an underlying pattern of the distribution of primes, it is possible to provide an even more compact viewpoint. According to a conjecture of Montgomery and Soundararajan [Montgomery and Soundararajan 2004], we should expect the primes in  $s_k$  to follow a normal distribution with mean

$$\mu_k = \frac{l_k}{\log p_{k+1}^2}$$

and standard deviation

$$\sigma_k = \frac{\sqrt{l_k(\log(p_{k+1}^2/l_k) + B)}}{\log p_{k+1}^2},$$

where  $B = 1 - \gamma - \log 2\pi$ . Actually,

$$\frac{\sqrt{l_k(\log(p_{k+1}^2/l_k) + B)}}{\log p_{k+1}^2} \sim \frac{\sqrt{l_k \log(p_{k+1}^2/l_k)}}{\log p_{k+1}^2},$$

but the latter expression does not give the same numerical accuracy for the data we have available, so we stick with  $\sigma_k$  as defined.

Next, we apply Montgomery and Soundararajan's conjecture to generate normalised versions of the prime counting functions  $\pi_k$ :

$$\bar{\pi}_k := \frac{\pi_k - \mu_k}{\sigma_k}.$$

Assuming the conjecture holds, the result of this step should be that  $\bar{\pi}_k$  is normally distributed with mean 0 and standard deviation 1. This expectation is accurately confirmed by Figure 4, where we (A) plot  $\bar{\pi}_k$  as a function of  $k$  and (B) display a histogram revealing the probability distribution of  $\bar{\pi}_k$ . As such, this result lends support to Montgomery and Soundararajan's conjecture. But more importantly, it promotes the fundamental idea that we can view  $\pi_k$  as a random variable with variance  $\mu_k$  and standard deviation  $\sigma_k$ .

We turn now to the global distribution of primes, where we want to examine the behaviour of the prime counting function  $\pi(x)$  when it is expressed in terms of the individual prime counting functions  $\pi_k$ :

$$\pi(p_{k+1}^2) = \sum_{j=1}^k \pi_j.$$

To best get an impression of how  $\pi(x)$  behaves, we plot—rather than  $\pi(x)$  itself—the error function

$$\epsilon(x) := \pi(x) - \text{li}(x),$$

which produces the result seen in Figure 5. The first observation we make is that the error function starts out being negative—a famed bias that continues for the full stretch of our data set. Nonetheless, as Littlewood proved [Littlewood 1914], if we persist in moving  $x$  towards infinity, the error term will eventually shift sign, and it will continue to do so infinitely many times over. When that happens for the first time, however, no one knows, though it has been proved that  $x = 1.39822 \times 10^{316}$  is

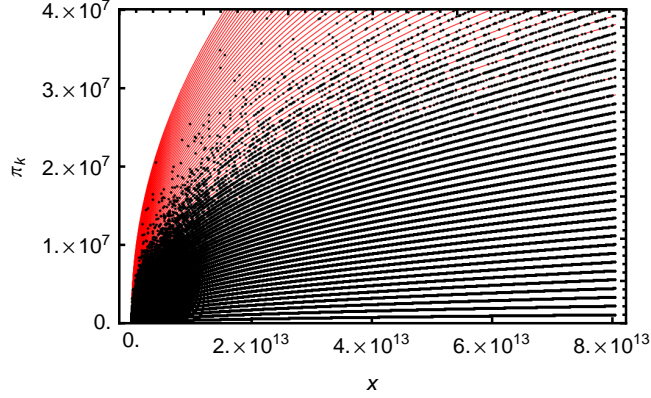


FIGURE 3. The number of primes  $\pi_k$  in each interval  $s_k$  plotted at the values  $x = p_{k+1}^2$ ,  $1 \leq k \leq 6 \times 10^5$  (black dots). The emerging curves relate to the specific gap values  $g$ ,  $g \in \{2, 4, \dots\}$ , where the bottom curve corresponds to the set of intervals  $s_k$  where  $g_k = 2$ , the second bottom  $g_k = 4$ , etc. The red curves show the corresponding theoretical estimates  $(2\sqrt{x}g - g^2)/\log x$ .

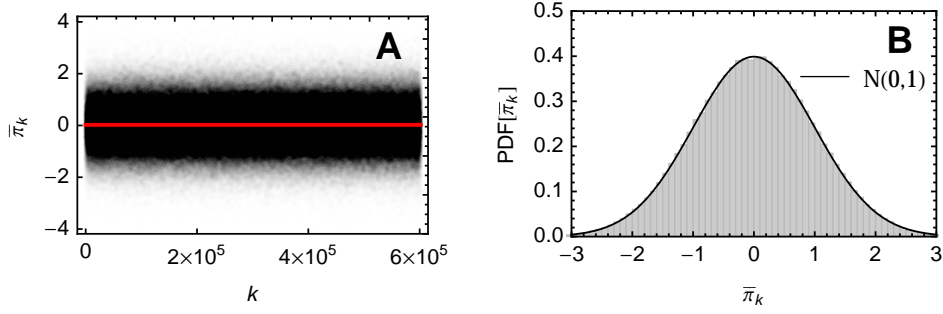


FIGURE 4. (A) The normalised prime counting functions  $\bar{\pi}_k$  plotted as a function of  $k$ ,  $1 \leq k \leq 6 \times 10^5$ . (B) A histogram displaying the numerical probability distribution of  $\bar{\pi}_k$ , overlaid with the PDF of the normal distribution  $N(0, 1)$  (black). The measured mean and standard deviation of  $\bar{\pi}_k$  are 0.000(7) and 1.000(8), respectively.

an upper bound for the event [Bays and Hudson 2000]. A probabilistic measure of the bias was provided in 1994, when Rubinstein and Sarnak [Rubinstein and Sarnak 1994]—assuming the Riemann hypothesis—calculated the logarithmic density of those  $x$  such that  $\pi(x) - \text{li}(x) > 0$  to be around 0.00000026.

According to Conjecture 1, however,  $\text{li}(x)$  is a continuum approximation to the discrete sum  $\sum_{j=1}^k l_j / \log p_{j+1}^2$ , and therefore we should expect the latter to be a better estimate of  $\pi(x)$ . As we can tell from Figure 5, where we also have included curves of  $\sum_{j=1}^k (l_j / \log p_j^2 - \text{li}_j)$  and  $\sum_{j=1}^k (l_j / \log p_{j+1}^2 - \text{li}_j)$ , the statistical evidence

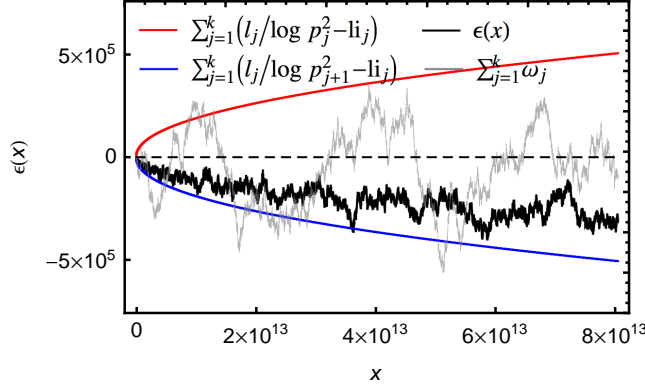


FIGURE 5. The error function  $\epsilon(x)$  plotted at the values  $x = p_{k+1}^2$ ,  $1 \leq k \leq 6 \times 10^5$  (black). Also shown are the differences  $\sum_{j=1}^k (l_j / \log p_j^2 - \text{li}_j)$  (red), and  $\sum_{j=1}^k (l_j / \log p_{j+1}^2 - \text{li}_j)$  (blue), as well as one realisation of the sum  $\sum_{j=1}^k \omega_j$  (grey), where the  $\omega_k$ s are independent random variables each satisfying  $N(0, \sigma_k)$ .

seems to indicate that the truth lies somewhere in between,

$$\sum_{j=1}^k \frac{l_j}{\log p_{j+1}^2} < E[\pi(p_{k+1}^2)] < \text{li}(p_{k+1}^2).$$

To more accurately interpret this observation, we normalise the curves in Figure 5 by

$$\Delta_k := \frac{1}{2} \sum_{j=1}^k \left( \frac{l_j}{\log p_j^2} - \frac{l_j}{\log p_{j+1}^2} \right),$$

which places the normalised differences

$$\left[ \sum_{j=1}^k \left( \frac{l_j}{\log p_{j+1}^2} - \text{li}_j \right) \right] / \Delta_k \quad \text{and} \quad \left[ \sum_{j=1}^k \left( \frac{l_j}{\log p_j^2} - \text{li}_j \right) \right] / \Delta_k$$

approximately at the constant lines -1 and 1 respectively. We also normalise the error function, defined by

$$E_k := \frac{\epsilon(p_{k+1}^2)}{\Delta_k}.$$

The resulting curves are shown in Figure 6A, plotted as functions of  $k$ . We note that on the scale of the data available,  $E_k$  fluctuates fairly stable around a mean value of  $-0.60$ . This becomes even more apparent if we display the data in histogram form, as done in Figure 6B, where we see that the empirical probability distribution of  $E_k$  resembles a normal distribution. A natural speculation therefore, is whether there in fact exists a constant  $c$  such that

$$E[\pi(p_{k+1}^2)] = c \cdot \text{li}(p_{k+1}^2) - (1 - c) \cdot \sum_{j=1}^k \frac{l_j}{\log p_{j+1}^2}.$$

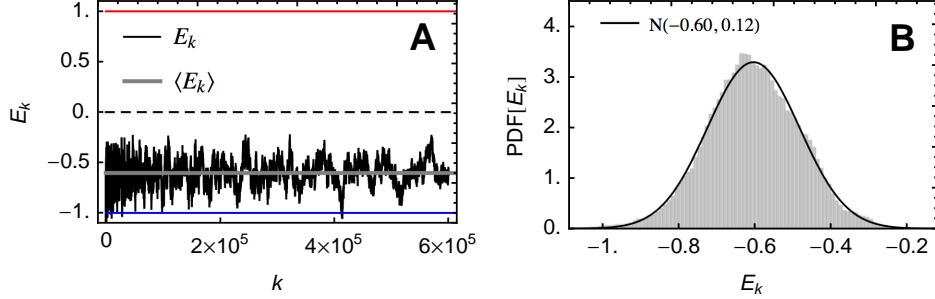


FIGURE 6. The normalised error function  $E_k$  plotted as a function of  $k$ ,  $1 \leq k \leq 6 \times 10^5$  (A), as well as a probability density histogram of all values of  $E_k$  (B). Also displayed in (B) is the PDF for a normal distribution with mean  $-0.60(4)$  and standard deviation  $0.12(1)$  estimated from all  $E_k$ .

Returning to Figure 5, we end with one final important observation: The error function  $\epsilon(x)$  does not conform to a sum of independent random variables. This is clearly visible when comparing with an example of an actual sum of independent random variables  $\omega_k$ , where  $\omega_k \sim N(0, \sigma_k)$ ; the error function does not display the same scale of fluctuations as  $\sum_k \omega_k$ , and as already discussed, lingers more densely around a mean value. As will be apparent later, what causes the discrepancy between the error function  $\epsilon(x)$  and  $\sum_{j=1}^k \omega_j$  are correlations between the prime counting functions  $\pi_k$ . These correlations even grow stronger for larger interval lengths  $l_k$ , possibly explaining why the normalised error function  $E_k$  in Figure 6A seems to fluctuate more slowly as  $k$  becomes large. Explaining these correlations and their effect on the error function will be the focus of the next sections. But for now, we summarise the numerical insights by the following remark:

**Remark 1.** The prime counting function  $\pi(x)$  behaves—and should be understood—as a sum of *correlated* random variables  $\pi_k$ , each normally distributed with mean  $\mu_k$  and standard deviation  $\sigma_k$  as conjectured in [Montgomery and Soundararajan 2004].

## 7. A RANDOM MODEL FOR THE DISTRIBUTION OF PRIMES

With the understanding we have developed so far, we now have in hand enough information to construct a random model of the prime counting function  $\pi(x)$ . We start with emphasising the fundamental observation underlying the model; the exact distribution of primes up to  $p_{k+1}^2$  is in its entirety build up of the  $k$  periodic sequences  $\rho_j(n)$ ,  $1 \leq j \leq k$ , each having the property that  $\rho_j(0) = p_j$ . We visualise this in the following table:

$n$	<b>0</b>	<b>1</b>	<b>2</b>	<b>3</b>	<b>4</b>	<b>5</b>	<b>6</b>	<b>7</b>	<b>8</b>	<b>9</b>	<b>10</b>	<b>11</b>	<b>12</b>	<b>13</b>	<b>14</b>	<b>15</b>	$\dots$
$\rho_1(n)$	<b><math>p_1</math></b>	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$p_1$	1	$\dots$
$\rho_2(n)$	<b><math>p_2</math></b>	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	1	1	$p_2$	$\dots$
$\rho_3(n)$	<b><math>p_3</math></b>	1	1	1	1	$p_3$	1	1	1	1	$p_3$	1	1	1	1	$p_3$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\rho_k(n)$	<b><math>p_k</math></b>	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	$\dots$

Importantly—and already discussed at length—within  $s_k$ , only the  $k$  first sequences

$\rho_j$ ,  $1 \leq j \leq k$ , are needed to completely determine the distribution of primes. Note also that the distribution of primes in  $s_k$  correlates with the distribution of primes in all previous intervals  $s_i$ ,  $1 \leq i < k$ , due to the periodicities of  $\rho_j$ ,  $1 \leq j \leq i$ .

In formulating the model—which we hereafter refer to as the *correlated random model*—the essential idea is to keep the intervals  $s_k$ , but to allow translations of the sequences  $\rho_j$ ,  $1 \leq j \leq k$ . By this approach, we stay close to the original structure of the primes, but will eventually be able to get a theoretical grip on the correlations between intervals; as becomes apparent, this is exactly what we need to give a proper explanation of the error function  $\epsilon(x)$ .

The central object in the model is the random variable  $\tilde{\Pi}(p_{k+1}^2)$ , which is defined in terms of a sum over the random variables  $\tilde{\Pi}_j$ :

$$\tilde{\Pi}(p_{k+1}^2) := \sum_{j=1}^k \tilde{\Pi}_j.$$

To generate the random variables  $\tilde{\Pi}_j$ ,  $1 \leq j \leq k$ , we randomly choose  $k$  corresponding integers  $m_j$ , where  $m_j \in \{0, 1, \dots, p_j - 1\}$  and then define the translated sequences

$$\hat{\rho}_j(n) := \rho_j(n + m_j).$$

As in Section 2, we construct a second arithmetic function; here with the purpose of locating all  $n$  such that  $n + m_i$ ,  $1 \leq i \leq j$ , is coprime to  $p_j$ :

$$\hat{R}_j(n) := \prod_{1 \leq i \leq j} \hat{\rho}_i(n).$$

The random variables  $\tilde{\Pi}_j$  are now obtained by

$$\tilde{\Pi}_j := |\{n : n \in s_j, \hat{R}_j(n) = 1\}|.$$

Since the sequences  $\hat{\rho}_j$ ,  $1 \leq j \leq k$ , are drawn independent from each other, the expectation value of  $\tilde{\Pi}_j$  is given by

$$\mathbb{E} [\tilde{\Pi}_j] = \tilde{\pi}_j = l_j \cdot \prod_{p \in \mathcal{P}_j} \left(1 - \frac{1}{p}\right) \sim 2e^{-\gamma} \frac{l_j}{\log p_{j+1}^2} \sim 2e^{-\gamma} \text{li}_j,$$

and correspondingly,

$$\mathbb{E} [\tilde{\Pi}(p_{k+1}^2)] = \sum_{j=1}^k \tilde{\pi}_j \sim 2e^{-\gamma} \sum_{j=1}^k \text{li}_j = 2e^{-\gamma} \text{li}(p_{k+1}^2).$$

Note that the model as stated does not produce the correct mean. This can easily be overcome by introducing the random variables

$$\Pi_j := \frac{e^\gamma}{2} \tilde{\Pi}_j \quad \text{and} \quad \Pi(p_{k+1}^2) := \sum_{j=1}^k \Pi_j.$$

However, our interest is not in the mean, but in the variance produced by this model, and for that purpose it is not essential whether the mean differs by a constant; we still expect the primes to be constrained by the same variance.

To continue, we write the variance of  $\tilde{\Pi}(p_{k+1}^2)$  as

$$\begin{aligned} \text{Var} \left[ \tilde{\Pi}(p_{k+1}^2) \right] &= \mathbb{E} \left[ \left( \sum_{j=1}^k (\tilde{\Pi}_j - \tilde{\pi}_j) \right)^2 \right] \\ &= \sum_{j=1}^k \mathbb{E} \left[ (\tilde{\Pi}_j - \tilde{\pi}_j)^2 \right] + 2 \sum_{1 \leq i < j \leq k} \mathbb{E} \left[ (\tilde{\Pi}_i - \tilde{\pi}_i) \cdot (\tilde{\Pi}_j - \tilde{\pi}_j) \right]. \end{aligned}$$

As we see, the first expression after the last equality is a sum over the individual variances, while the second expression contains all covariance terms; they are all appearing twice, thereby explaining the factor 2. Crucially, we can express the variance theoretically, and will do so in the next section.

For comparison, we also consider what we will refer to as the *uncorrelated random model*, where the sequences  $\hat{\rho}_j$ ,  $1 \leq j \leq k$ , are drawn anew for each interval  $s_k$ . We formulate this model in terms of the random variables

$$\hat{\Pi}_j, 1 \leq j \leq k, \quad \text{and} \quad \hat{\Pi}(p_{k+1}^2) := \sum_{j=1}^k \hat{\Pi}_j.$$

The expectation values of these random variables are the same as above for the correlated random model, but now the covariances between intervals are zero, so that the variance is simply

$$\text{Var} \left[ \hat{\Pi}(p_{k+1}^2) \right] = \sum_{j=1}^k \mathbb{E} \left[ (\hat{\Pi}_j - \tilde{\pi}_j)^2 \right].$$

To illustrate the behaviour of the random models, we generate 275 realisations of  $\tilde{\Pi}(p_{k+1}^2)$ , as well as 275 realisations of  $\hat{\Pi}(p_{k+1}^2)$ , for  $1 \leq k \leq 10^4$ . The result is displayed in Figure 7, where the different realisations are plotted with their mean subtracted. Notable is that the correlated random model exhibits less variance than the uncorrelated model, indicating that the covariance terms in the correlated model in general are negative. In the same figure, we also plot the error function

$$\epsilon(p_{k+1}^2) := \pi(p_{k+1}^2) - \text{li}(p_{k+1}^2)$$

as well as the adjusted error function

$$\epsilon_0(p_{k+1}^2) := \pi(p_{k+1}^2) - \left[ c \cdot \text{li}(p_{k+1}^2) - (1-c) \cdot \sum_{j=1}^k \frac{l_j}{\log p_{j+1}^2} \right],$$

where  $c \approx 1 - 0.604$ . It is evident that both of these error functions display a behaviour that is closer to that of the correlated model.

It is worth contrasting the correlated random model with the famous Cramer random model [Cramér 1936] (see also [Granville 1995a] for a thorough review of Cramer's work). In Cramer's model each integer  $n$ ,  $1 \leq n \leq x$ , has the probability  $1/\log n$  of being prime, which gives a correct estimate of the number of primes across long enough intervals. While the model has proved very valuable in conjecturing different properties of the prime numbers, for example on the distribution of prime gaps, it also has the apparent weakness that it does not preserve certain central characteristics of the primes, such as the fact that an even number other than 2 can never be prime. In our suggested model, on the other hand, we move away from



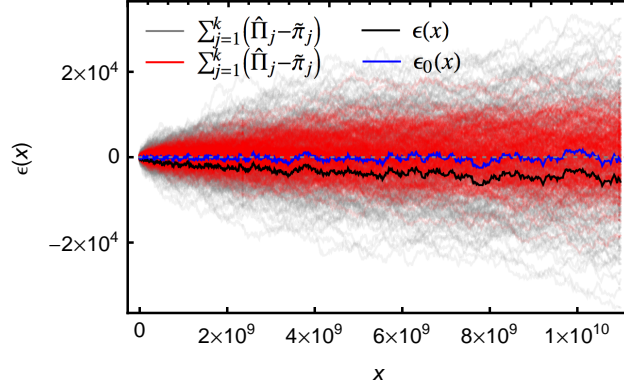


FIGURE 7. The error function  $\epsilon(x)$  (black) and the adjusted error function  $\epsilon_0(x)$  (blue) plotted at the values  $x = p_{k+1}^2$ ,  $1 \leq k \leq 10^4$ , together with 275 realisations of the correlated random model (red) and 275 realisations of the uncorrelated random model (grey).

looking at each integer. Rather, the number of primes in each interval  $s_k$  is treated as a random variable, and the correlation structure between intervals emerging from the underlying periodicities of  $\rho_k$  is kept intact. In this sense, we expect our choice of model to better describe the behaviour of the prime counting function  $\pi(x)$ .

#### 8. THEORETICAL EXPRESSIONS FOR THE VARIANCE OF $\tilde{\Pi}(p_{k+1}^2)$

In an elementary calculation, Hausman and Shapiro [Hausman and Shapiro 1973] derived the variance of the number of reduced residues modulo  $n$  in an arbitrary interval of length  $h$ . We here extend to the general case of the covariance between two intervals  $A_1$  and  $A_2$  with corresponding lengths  $h_1$  and  $h_2$  that are separated by a distance  $q$ , and where in  $A_1$  we regard the number of reduced residues modulo  $n_1$ , and in  $A_2$  the number of reduced residues modulo  $n_2$ . The resulting expressions for the covariance we present here follows from a straight forward adaption of Hausman and Shapiro's proof.

Let

$$f_n(m) := \begin{cases} 1 & \text{if } (m, n) = 1, \\ 0 & \text{if } (m, n) > 1. \end{cases}$$

The function of  $f_n(m)$  is to weed out coprimes to  $n$ . To count the number of such in an interval of length  $h$ , we define

$$F_n(m, h) := \sum_{r=m}^{m+h-1} f_n(r).$$

From a probabilistic perspective, the expected number of coprimes to  $n$  in an interval of length  $h$  is given by

$$h \frac{\Phi(n)}{n},$$

where  $\Phi(n)$  is Euler's totient function.

Denote now the number of coprimes to  $n_1$  in  $A_1$  and the number of coprimes to  $n_2$  in  $A_2$  by  $F_{n_1}(m, h_1)$  and  $F_{n_2}(m + q, h_2)$ , respectively, where we assume  $n_1 \leq n_2$  and  $q \geq 1$ . Then the covariance between  $F_{n_1}(m, h_1)$  and  $F_{n_2}(m + q, h_2)$  is given by

$$\begin{aligned}
 G(n_1, n_2, h_1, h_2, q) \\
 (9) \quad &:= \frac{1}{n_1 n_2} \sum_{m=1}^{n_1 n_2} \left( F_{n_1}(m, h_1) - h_1 \frac{\Phi(n_1)}{n_1} \right) \left( F_{n_2}(m + q, h_2) - h_2 \frac{\Phi(n_2)}{n_2} \right) \\
 &= \frac{1}{n_1 n_2} \sum_{m=1}^{n_1 n_2} F_{n_1}(m, h_1) F_{n_2}(m + q, h_2) - h_1 h_2 \frac{\Phi(n_1 n_2)}{n_1 n_2}.
 \end{aligned}$$

Note that  $G$  is periodic with respect to  $q$  and has periodicity  $n_1$ ,

$$G(n_1, n_2, h_1, h_2, q + n_1) = G(n_1, n_2, h_1, h_2, q).$$

For brevity, we only continue with the case when  $h_2 > h_1$  and  $n_1$  and  $n_2$  are both even. Similar expressions can be derived in the other cases. Under this assumption, and introducing the abbreviation

$$Q(n_1, n_2) := \frac{1}{2} n_1 n_2 \prod_{p | (n_1 n_2 / (n_1, n_2)^2)} \left( 1 - \frac{1}{p} \right) \prod_{p | (n_1, n_2)'} \left( 1 - \frac{2}{p} \right),$$

where  $n'$  denotes the largest odd divisor of  $n$ , we can express the sum after the last equality in (9) as

$$\begin{aligned}
 &\sum_{m=1}^{n_1 n_2} F_{n_1}(m, h_1) F_{n_2}(m + q, h_2) = \\
 &Q(n_1, n_2) \sum_{\substack{k=1 \\ 2 | q+k-1}}^{h_2-h_1+1} h_1 \prod_{\substack{p | (n_1, n_2)' \\ p | (q+k-1)}} \left( 1 + \frac{1}{p-2} \right) + \\
 &Q(n_1, n_2) \sum_{\substack{k=1 \\ 2 | (q+h_2-h_1+k)}}^{h_1-1} (h_1 - k) \prod_{\substack{p | (n_1, n_2)' \\ p | (q+h_2-h_1+k)}} \left( 1 + \frac{1}{p-2} \right) + \\
 &Q(n_1, n_2) \sum_{\substack{k=1 \\ 2 | (q-k)}}^{h_1-1} (h_1 - k) \prod_{\substack{p | (n_1, n_2)' \\ p | (q-k)}} \left( 1 + \frac{1}{p-2} \right).
 \end{aligned}$$

For computational purposes,  $G(n_1, n_2, h_1, h_2, q)$  is most efficiently calculated using this expression, and it underlies later numerical examples. But it is possible to derive a more compact theoretical expression in terms of the Möbius function [see (7)]. We also need the following definitions: For any square-free integer  $n$ ,

$$\rho(n) := \prod_{p | n} (p - 2);$$

$$\gamma^+(h, q, d) := \begin{cases} 0 & \text{if } d \nmid q + i \text{ for all } i, \text{ where } 1 \leq i \leq d \left\{ \frac{h}{d} \right\}, \\ 1 & \text{if } d \mid q + i \text{ for some } i, \text{ where } 1 \leq i \leq d \left\{ \frac{h}{d} \right\}; \end{cases}$$

$$\gamma^-(h, q, d) := \begin{cases} 0 & \text{if } d \nmid q - i \text{ for all } i, \text{ where } 1 \leq i \leq d \left\{ \frac{h}{d} \right\}, \\ 1 & \text{if } d \mid q - i \text{ for some } i, \text{ where } 1 \leq i \leq d \left\{ \frac{h}{d} \right\}; \end{cases}$$

and

$$\left\{ \frac{q}{d} \right\}_1 := \begin{cases} \left\{ \frac{q}{d} \right\} & \text{if } d \nmid q, \\ 1 & \text{if } d \mid q. \end{cases}$$

Then we can write the covariance  $G(n_1, n_2, h_1, h_2, q)$  on the equivalent form

$$(10) \quad G(n_1, n_2, h_1, h_2, q) = \frac{1}{n_1 n_2} Q(n_1, n_2) \sum_{d|(n_1, n_2)'} \frac{\mu^2(d)}{\rho(d)} \mathcal{B}(h_1, h_2, q, d),$$

where

$$\begin{aligned} & \mathcal{B}(h_1, h_2, q, d) \\ &= h_1 \left( \frac{1}{2d} - \left\{ \frac{h_2 - h_1 + 1}{2d} \right\} + \left\{ \frac{q + h_2 - h_1}{2d} \right\} - \left\{ \frac{q}{2d} \right\}_1 \right) \\ & - 2d \left\{ \frac{h_1}{2d} \right\} \left( \left\{ \frac{h_1}{2d} \right\} + \left\{ \frac{q + h_2 - h_1}{2d} \right\} - \left\{ \frac{q}{2d} \right\}_1 \right) \\ & + \gamma^+(h_2 - h_1 + 1, q - 1, 2d) h_1 \\ & + \gamma^+(h_1, q + h_2 - h_1, 2d) 2d \left( \left\{ \frac{h_1}{2d} \right\} - \left( 1 - \left\{ \frac{q + h_2 - h_1}{2d} \right\} \right) \right) \\ & + \gamma^-(h_1, q, 2d) 2d \left( \left\{ \frac{h_1}{2d} \right\} - \left\{ \frac{q}{2d} \right\}_1 \right). \end{aligned}$$

Numerical investigations hints to a possible simplification of this expression, as some of the terms appear to always cancel each other, but this remains to be shown theoretically.

Whenever  $A_1 = A_2$ , so that  $h_1 = h_2 = h$  and  $n_1 = n_2 = n$ , we can write  $H(n, h) := G(n, n, h, h, 0)$ , and (10) reduces to the variance derived in [Hausman and Shapiro 1973],

$$H(n, h) = \prod_{p|n'} \left( 1 - \frac{2}{p} \right) \sum_{d|n'} \frac{\mu^2(d)}{\rho(d)} d \left\{ \frac{h}{2d} \right\} \left( 1 - \left\{ \frac{h}{2d} \right\} \right).$$

Note that if we apply the inequality  $\{x\}(1 - \{x\}) \leq x$ , we obtain the upper bound

$$(11) \quad H(n, h) \leq h \frac{\Phi(n)}{n}.$$

Returning to the random models in Section 7, we can now express the variance of the correlated model as

$$\begin{aligned} \text{Var} \left[ \tilde{\Pi}(p_{k+1}^2) \right] &= \sum_{j=1}^k \mathbb{E} \left[ \left( \tilde{\Pi}_j - \tilde{\pi}_j \right)^2 \right] + 2 \sum_{1 \leq i < j \leq k} \mathbb{E} \left[ \left( \tilde{\Pi}_i - \tilde{\pi}_i \right) \cdot \left( \tilde{\Pi}_j - \tilde{\pi}_j \right) \right] \\ &= \sum_{j=1}^k H(p_j \#, l_j) + 2 \sum_{1 \leq i < j \leq k} G(p_i \#, p_j \#, l_i, l_j, p_{j+1}^2 - p_{i+1}^2). \end{aligned}$$

Likewise, for the uncorrelated model,

$$\text{Var} \left[ \hat{\Pi}(p_{k+1}^2) \right] = \sum_{j=1}^k \mathbb{E} \left[ \left( \hat{\Pi}_j - \tilde{\pi}_j \right)^2 \right] = \sum_{j=1}^k H(p_j \#, l_j).$$

From (11) it follows that

$$H(p_j \#, l_j) \leq l_j \cdot \frac{\Phi(p_j \#)}{p_j \#} = l_j \cdot \prod_{p \in \mathcal{P}_j} \left( 1 - \frac{1}{p} \right) \sim 2e^{-\gamma} \frac{l_j}{\log p_{j+1}^2} \sim 2e^{-\gamma} \text{li}_j,$$

which results in the variance of the uncorrelated model having the upper bound

$$\text{Var} \left[ \hat{\Pi}(p_{k+1}^2) \right] \leq 2e^{-\gamma} \sum_{j=1}^k \text{li}_j = 2e^{-\gamma} \text{li}(p_{k+1}^2).$$

## 9. PROPERTIES OF THE COVARIANCE FUNCTION

In this section we present some general considerations on the behaviour of the covariance function  $G(n_1, n_2, h_1, h_2, q)$ . Let us first have a look at the role of the relative position  $q$  of two intervals, when both are assumed to be sieved by the  $k$  first primes. As an example of this scenario, we show in Figure 8A a plot of  $G(p_{100} \#, p_{100} \#, h_1, 2h_1, q)$  as a function of  $q$  for different values of  $h_1$ . We observe that in each case, the covariance function reaches a negative minimum value when  $q = h_1$ —the exact situation when the two intervals are adjacent—before slowly climbing towards zero again. In addition, the minimum values grow in size as the length of the intervals increases. While not shown in Figure 8A, this result also holds for arbitrary values of  $h_2$ . Note, however, that the observed minimum values are not unique, as shown in Figure 8B, where we have plotted  $G(p_k \#, p_k \#, h_1, h_1, q)$  across its full period  $p_k \#$  for two values of  $h_1$ .

The reason why  $q = h_1$  corresponds to a minimum value of the covariance function can be understood in terms of the following heuristic: For the sake of argument, assume that  $A_1$  and  $A_2$  are both sieved by  $p_k$  only, and that their lengths  $h_1$  and  $h_2$  satisfy  $h_1 + h_2 \leq p_k$ . Then we have that

$$F_{p_k}(m, h_1) = h_1 - 1 \quad \text{or} \quad F_{p_k}(m, h_1) = h_1,$$

depending respectively on whether the interval  $A_1$  lies across a multiple of  $p_k$  or not. Likewise in the case of  $A_2$ :

$$F_{p_k}(m, h_2) = h_2 - 1, \quad \text{or} \quad F_{p_k}(m, h_2) = h_2.$$

Under these assumptions, we have that

$$G(p_k, p_k, h_1, h_2, q) = \frac{1}{p_k} \sum_{m=1}^{p_k} F_{p_k}(m, h_1) F_{p_k}(m + q, h_2) - h_1 h_2 \left( \frac{p_k - 1}{p_k} \right)^2.$$

When  $A_1$  and  $A_2$  do not overlap, that is,  $q \geq h_1$ , we can write the sum after the equality in terms of three contributions,

$$\begin{aligned} & \sum_{m=1}^{p_k} F_{p_k}(m, h_1) F_{p_k}(m + q, h_2) \\ &= (p_k - h_1 - h_2) \cdot [h_1 h_2] + h_1 \cdot [h_2(h_1 - 1)] + h_2 \cdot [h_1(h_2 - 1)] \\ &= h_1 h_2 (p_k - 2). \end{aligned}$$

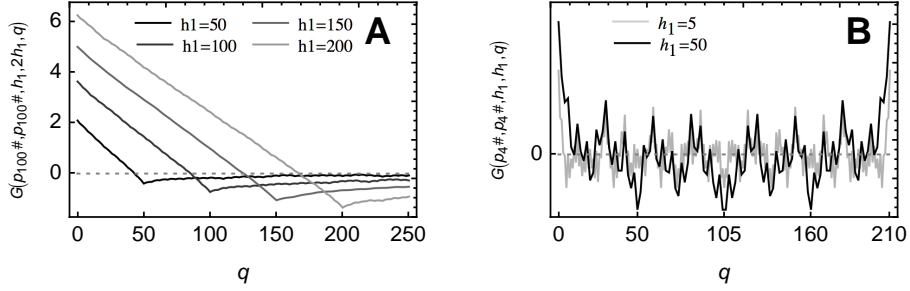


FIGURE 8. Covariance plots: (A)  $G(p_{100}\#, p_{100}\#, h_1, 2h_1, q)$  plotted for four different values of  $h_1$  as a function of  $q$ . In each case, the covariance function is minimum at  $q = h_1$ . (B)  $G(p_4\#, p_4\#, h_1, h_1, q)$  plotted for two different values of  $h_1$  as a function of  $q$  across its full period. The plot illustrates that the minimum found at  $q = h_1$  is in general not unique.

And if  $A_1$  and  $A_2$  do overlap, that is,  $0 \leq q < h_1$ , in terms of four contributions,

$$\begin{aligned}
 & \sum_{m=1}^{p_k} F_{p_k}(m, h_1) F_{p_k}(m+q, h_2) \\
 &= (p_k - h_2 - q) \cdot [h_1 h_2] + q \cdot [(h_1 - 1)h_2] + (h_1 - q) \cdot [(h_1 - 1)(h_2 - 1)] \\
 &\quad + (h_2 - h_1 + q) \cdot [h_1(h_2 - 1)] \\
 &= -q + h_1 + h_1 h_2 (p_k - 2).
 \end{aligned}$$

Comparing these two sums, we see that  $G(p_k, p_k, h_1, h_2, q)$  is minimised whenever  $q \geq h_1$ , in which case

$$G(p_k, p_k, h_1, h_2, q) = -\frac{h_1 h_2}{p_k^2} < 0.$$

Note that the larger the product  $h_1 h_2$  is, the larger the absolute value of the covariance at its minimum, as we observe in Figure 8A.

Letting go of the restriction that  $h_1 + h_2 \leq p_k$  complicates the heuristic, since more careful accounting is needed; as does sieving the intervals by multiple primes. It seems realistic, however, that a proof can be produced to show that in general,  $G(p_k\#, p_k\#, h_1, h_2, q)$  takes on a negative minimum value whenever  $q = h_1$ .

Aside the relative positions of intervals, a second aspect to consider is what primes each interval is sieved by. Again, assume two intervals, now with lengths  $h_1$  and  $h_2$ , where the first is sieved by the  $i$  first primes, and the second by the  $j$  first primes. Then we can write the covariance function as

$$G(p_i\#, p_j\#, h_1, h_2, q) = c_1 \prod_{p|(p_j\#/p_i\#)} \left(1 - \frac{1}{p}\right) - c_2 \prod_{p \in \mathcal{P}_j} \left(1 - \frac{1}{p}\right),$$

where  $c_1$  and  $c_2$  are constants. Both products approaches 0 as  $j \rightarrow \infty$ , and therefore we have that

$$\lim_{j \rightarrow \infty} G(p_i\#, p_j\#, h_1, h_2, q) = 0.$$

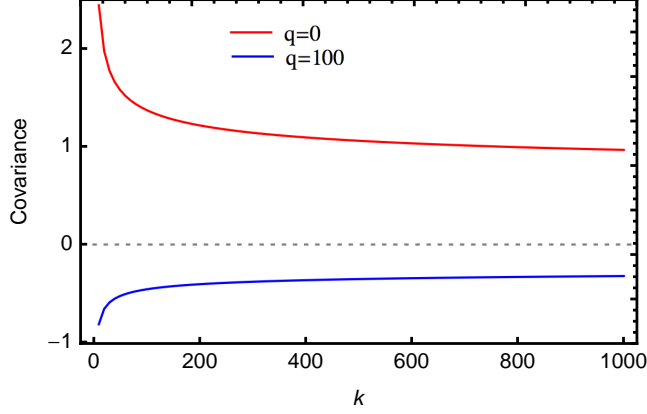


FIGURE 9. The covariance function  $G(p_{10}\#, p_k\#, 100, 100, q)$  plotted for two different values of  $q$  as a function of  $10 \leq k \leq 1000$ .

This limit behaviour is illustrated in Figure 9;  $G(p_i\#, p_j\#, h_1, h_2, q)$  is plotted for two values of  $q$ , and in both cases it moves slowly towards 0 as  $j$  increases.

Combining these two insights—the covariance function is minimised for adjacent intervals and approaches zero as the difference in sieving steps between intervals increases—we anticipate that  $G(p_i\#, p_j\#, l_i, l_j, p_{j+1}^2 - p_{i+1}^2)$  is negative when  $j - i$  is small, increasingly so when  $i$  and  $j$  both become large, and grows smaller in absolute value as the difference  $j - i$  increases. Hence, the sum

$$\sum_{1 \leq i < j \leq k} G(p_i\#, p_j\#, l_i, l_j, p_{j+1}^2 - p_{i+1}^2)$$

should be dominated by negative terms, implying that the variance of the correlated random model is bounded by the same upper bound as the uncorrelated random model. Thus,

$$\text{Var} \left[ \tilde{\Pi}(p_{k+1}^2) \right] \leq \sum_{i=1}^k H(p_i\#, l_i) \leq 2e^{-\gamma} \text{li}(p_{k+1}^2).$$

## 10. COMPARING RANDOM MODELS AND EMPIRICAL DATA

In order to verify the predictions stated in the previous section we now consider four cases: 1) The theoretical variance of the correlated random model; 2) The variance of the correlated random model estimated from sampled realisations of  $\tilde{\Pi}(p_{k+1}^2)$ ; 3) The estimated variance of the prime counting function  $\pi(x)$  under the assumption that  $E[\pi(x)] = \text{li}(x)$ ; and 4) The estimated variance of the prime counting function  $\pi(x)$  under the assumption that

$$E[\pi(p_{k+1}^2)] = c \cdot \text{li}(p_{k+1}^2) - (1 - c) \cdot \sum_{i=1}^k l_i / \log p_{i+1}^2.$$

Let us start with the separate contribution from the covariance terms between intervals. In the theoretical case, we write the sum

$$\sum_{1 \leq i < j \leq k} G(p_i \#, p_j \#, l_i, l_j, p_{j+1}^2 - p_{i+1}^2)$$

on the equivalent form

$$\sum_{j=1}^{k-1} \left( \sum_{i=1}^{k-j} G(p_i \#, p_{i+j} \#, l_i, l_{i+j}, p_{i+j+1}^2 - p_{i+1}^2) \right),$$

and denote the expression in parentheses by

$$\kappa_{\text{th}}(j) := \sum_{i=1}^{k-j} G(p_i \#, p_{i+j} \#, l_i, l_{i+j}, p_{i+j+1}^2 - p_{i+1}^2),$$

where the label "th" is short for theory.  $\kappa_{\text{th}}(j)$  is therefore the sum over all covariance terms originating from  $j$ th nearest intervals. For example,  $\kappa_{\text{th}}(1)$  gives the sum over all covariance terms for neighbouring intervals. We also need the total remaining contribution to the covariance for all  $j$  larger than some value  $d$ , which we obtain by

$$\sum_{j>d}^{k-1} \kappa_{\text{th}}(j).$$

Now to the second case, where we draw samples of  $\tilde{\Pi}(p_{k+1}^2) = \sum_{i=1}^k \tilde{\Pi}_i$ . We define the error function in each interval  $j$  as  $\epsilon_{\text{sim},j} := \tilde{\Pi}_j - \tilde{\pi}_j$ , where the label "sim" is short for simulation, and write the global error function as

$$\epsilon_{\text{sim}}(p_{k+1}^2) := \sum_{j=1}^k (\tilde{\Pi}_j - \tilde{\pi}_j) = \sum_{j=1}^k \epsilon_{\text{sim},j}.$$

Then the measured variance of one realisation of  $\tilde{\Pi}(p_{k+1}^2)$  can be expressed as

$$\text{Var} \left[ \tilde{\Pi}(p_{k+1}^2) \right] = \left( \sum_{j=1}^k \epsilon_{\text{sim},j} \right)^2 = \sum_{j=1}^k \epsilon_{\text{sim},j}^2 + 2 \sum_{1 \leq i < j \leq k} \epsilon_{\text{sim},i} \cdot \epsilon_{\text{sim},j}.$$

Similar to the theoretical case, we rewrite the sum

$$\sum_{1 \leq i < j \leq k} \epsilon_{\text{sim},i} \cdot \epsilon_{\text{sim},j}$$

as

$$\sum_{j=1}^{k-1} \left( \sum_{i=1}^{k-j} \epsilon_{\text{sim},i} \cdot \epsilon_{\text{sim},i+j} \right).$$

Since we are interested in the average taken over many realisations, we define

$$\kappa_{\text{sim}}(j) := \left\langle \sum_{i=1}^{k-j} \epsilon_{\text{sim},i} \cdot \epsilon_{\text{sim},i+j} \right\rangle,$$

and the total remaining contribution to the covariance for all  $j$  larger than some value  $d$ , by

$$\sum_{j>d}^k \kappa_{\text{sim}}(j).$$

Finally, in the cases of the primes, we define

$$\epsilon_{pr,j} := \pi_j - \text{li}_j \quad \text{and} \quad \epsilon_{ad,j} := \pi_j - \left[ c \cdot \text{li}_j - (1-c) \cdot \frac{l_j}{\log p_{j+1}^2} \right],$$

where "pr" and "ad" denote primes and primes with adjusted mean, and obtain

$$\epsilon_{pr}(p_{k+1}^2) := \sum_{j=1}^k \epsilon_{pr,j} \quad \text{and} \quad \epsilon_{ad}(p_{k+1}^2) := \sum_{j=1}^k \epsilon_{ad,j}.$$

Then, entirely analogous to the previous cases, we obtain

$$\kappa_{pr}(j) := \sum_{i=1}^{k-j} \epsilon_{pr,i} \cdot \epsilon_{pr,i+j} \quad \text{and} \quad \sum_{j>d}^k \kappa_{pr}(j),$$

and

$$\kappa_{ad}(j) := \sum_{i=1}^{k-j} \epsilon_{ad,i} \cdot \epsilon_{ad,i+j} \quad \text{and} \quad \sum_{j>d}^k \kappa_{ad}(j).$$

We now plot  $\kappa_{\text{th}}(j)$ ,  $\kappa_{\text{sim}}(j)$ ,  $\kappa_{pr}(j)$ , and  $\kappa_{ad}(j)$  for different values of  $j$  up to a maximum  $d$ , as well as the remainder terms, all shown in Figure 10. The different plots are across different ranges of  $k$ —limited to smaller  $k$  in the first two cases due to increasing computational cost for large  $k$ —but we observe that they are all qualitatively similar, confirming that the primes behave according to the correlated random model. For direct comparison, we include  $\kappa_{pr}(1)$  and  $\kappa_{pr}(2)$  also in the plots of  $\kappa_{\text{th}}(j)$  and  $\kappa_{\text{sim}}(j)$ . At this short range, however, the fluctuations of  $\kappa_{pr}(1)$  and  $\kappa_{pr}(2)$  are rather significant, so we are not able to determine whether  $\kappa_{pr}(j)$  converges to the theoretical value of the correlated random model in the limit of large  $k$ . In fact, as will be evident when we plot the total variance in each case, it appears that the covariance functions  $\kappa_{pr}(j)$  and  $\kappa_{ad}(j)$  in general lie below  $\kappa_{\text{th}}(j)$  as  $k$  grows large, and this might be what we actually observe in Figure 10A and more convincingly in Figure 10B.

As we anticipated in the previous section, the contribution to the covariance is largest when  $j$  is small;  $\kappa_a(1)$  is more than a factor two larger than  $\kappa_a(2)$  (here "a" denotes any of "th", "sim", "pr", or "ad"). And as  $j$  grows larger,  $\kappa_a(j)$  approaches zero in absolute value. Note though that while we can observe some  $\kappa_a(j)$  being positive, the remainder terms in all cases add up to negative values.

Comparing Figure 10C and Figure 10D, we see that for small  $j$ ,  $\kappa_{pr}(j)$  and  $\kappa_{ad}(j)$  are almost identical. However, for  $j$  large the remainder terms differ significantly. Assuming  $\text{li}(x)$  to be the correct expectation value of  $\pi(x)$  produces large fluctuations in the remainder term, while a much smoother curve results from assuming the adjusted expectation value.



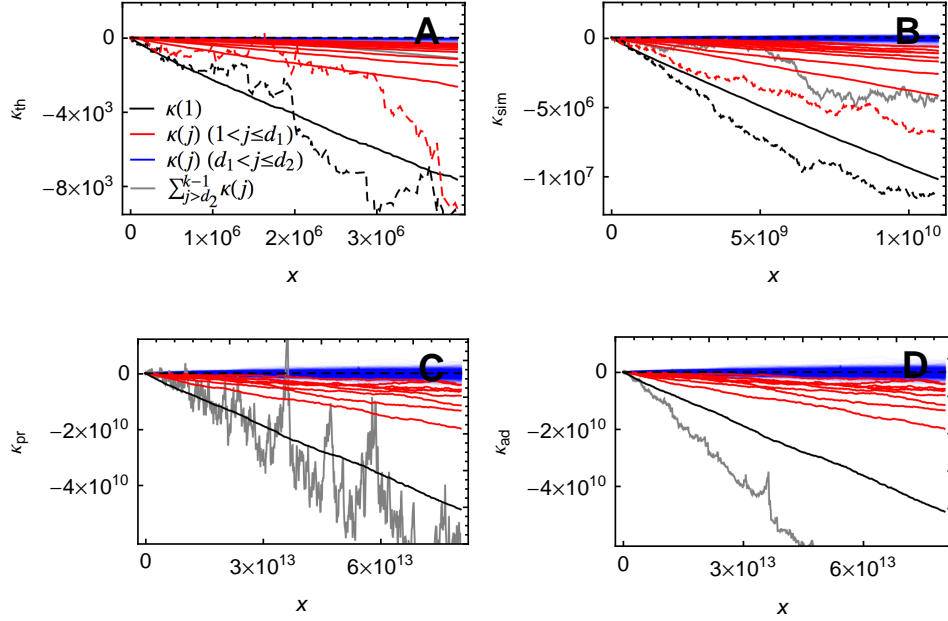


FIGURE 10. The covariance sums  $\kappa_{th}(j)$  (A),  $\kappa_{sim}(j)$  (B),  $\kappa_{pr}(j)$  (C), and  $\kappa_{ad}(j)$  (D) plotted at the values  $x = p_{k+1}^2$ . The legend applies for all three covariance sums. In (A),  $1 \leq k \leq 300$ , and data are plotted for each value of  $k$ ,  $d_1 = 10$ , and  $d_2 = 50$ ; In (B),  $1 \leq k \leq 10^3$ , and data are plotted for every 10th value of  $k$ ,  $d_1 = 10$ , and  $d_2 = 200$ . The curves are obtained by averaging over 275 individual samples; In (C) and (D),  $1 \leq k \leq 6 \times 10^5$ , and data are plotted for every 500th value of  $k$ ,  $d_1 = 10$ , and  $d_2 = 1000$ . In (A) and (B),  $\kappa_{pr}(1)$  and  $\kappa_{pr}(2)$  are included for comparison (dashed black and red, respectively).

Next, we continue by examining the total variance in each case, or rather, the standard deviations, that are given by

$$\begin{aligned}\sigma_{th}(p_{k+1}^2) &:= \sqrt{\text{Var}[\tilde{\Pi}(p_{k+1}^2)]}, \\ \sigma_{sim}(p_{k+1}^2) &:= \sqrt{\left\langle \left( \sum_{i=1}^k \epsilon_{sim,j} \right)^2 \right\rangle}, \\ \sigma_{pr}(p_{k+1}^2) &:= \sqrt{\left( \sum_{j=1}^k \epsilon_{pr,j} \right)^2}, \quad \text{and} \\ \sigma_{ad}(p_{k+1}^2) &:= \sqrt{\left( \sum_{j=1}^k \epsilon_{ad,j} \right)^2}.\end{aligned}$$

In addition, we consider the standard deviations under the assumption that all covariance terms are zero, that is, corresponding to the uncorrelated random model. Hence,

$$\begin{aligned}\sigma_{\text{th},0}(p_{k+1}^2) &:= \sqrt{\sum_{j=1}^k H(p_j\#, l_j)}, \\ \sigma_{\text{sim},0}(p_{k+1}^2) &:= \sqrt{\left\langle \sum_{j=1}^k \epsilon_{\text{sim},j}^2 \right\rangle}, \\ \sigma_{\text{pr},0}(p_{k+1}^2) &:= \sqrt{\sum_{j=1}^k \epsilon_{\text{pr},j}^2}, \quad \text{and} \\ \sigma_{\text{ad},0}(p_{k+1}^2) &:= \sqrt{\sum_{j=1}^k \epsilon_{\text{ad},j}^2}.\end{aligned}$$

Finally, we include the upper bound we found earlier for the standard deviation of the uncorrelated random model:

$$\sigma_{\text{ub}}(p_{k+1}^2) := \sqrt{2e^{-\gamma} \text{li}(p_{k+1}^2)}.$$

The different standard deviations are plotted in Figure 11. Under the assumption of the uncorrelated model, we observe that the curves of  $\sigma_{\text{pr},0}$  and  $\sigma_{\text{ad},0}$  lie on top of each other and are not visibly different. They also lie close to the theoretical value  $\sigma_{\text{th},0}$ , as shown in Figure 11A, and in fact seems to converge to  $\sigma_{\text{sim},0}$ , as evidenced in Figure 11B.

Including correlations, we note in Figure 11A that  $\sigma_{\text{pr}}$  starts out larger than  $\sigma_{\text{th}}$ , but seems to stabilise below  $\sigma_{\text{sim}}$  in Figure 11B. It appears that this trends continues in Figure 11C, but without comparison with the correlated random model, we cannot state this with certainty. In all cases,  $\sigma_{\text{ad}}$  is small compared to  $\sigma_{\text{pr}}$ .

As expected from the covariance contribution being negative, the total variance is always smaller than the variance of the uncorrelated model, which again is bounded by  $\sigma_{\text{ub}}$ , placed far above any of the other curves.

## 11. THE DISTRIBUTION OF PRIMES AND THE RIEMANN HYPOTHESIS

Considered an outstanding challenge in number theory, the Riemann hypothesis states that the real part of every non-trivial zero of the Riemann zeta function is  $1/2$ , where the Riemann zeta function is defined by

$$\zeta(s) := \sum_{n=1}^{\infty} \frac{1}{n^s}.$$

The Riemann zeta function ties in tightly with the distribution of primes through the Euler product formula, which states that

$$\zeta(s) = \prod_{p \in \mathcal{P}} \left(1 - \frac{1}{p^s}\right)^{-1}.$$

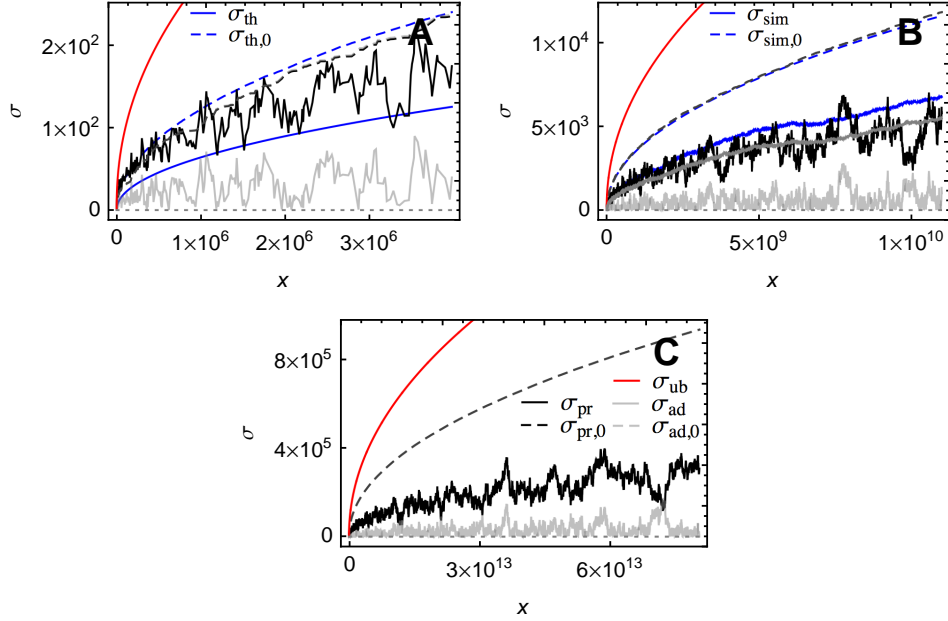


FIGURE 11. Different standard variance functions plotted at the values  $x = p_{k+1}^2$  (calculated from theory, samples from the correlated random model, and from the prime counting function  $\pi(x)$ ). The legends in (C) also applies to (A) and (B). In (A),  $1 \leq k \leq 300$ , and data are plotted for each value of  $k$ ; In (B),  $1 \leq k \leq 10^3$ , and data are plotted for every 10th value of  $k$ . The curves for  $\sigma_{sim}$  and  $\sigma_{sim,0}$  are obtained by averaging over 275 individual samples; In (C),  $1 \leq k \leq 6 \times 10^5$ , and data are plotted for every 100th value of  $k$ .

Our particular interest is in the fact that the Riemann hypothesis is equivalent to the upper bound on the error function being

$$|\pi(x) - \text{li}(x)| = O(\sqrt{x} \log x),$$

a statement proven by von Koch in 1901 [von Koch 1901].

Koch's criteria is far from being a tight upper bound; it is much wider than what we have conjectured—and justified—in the previous sections, namely that for  $x$  large enough,

$$(12) \quad |\pi(x) - \text{li}(x)| < 2e^{-\gamma} \sqrt{\text{li}(x)}.$$

In fact, we know enough to conjecture an even tighter upper bound, expressed discretely at the values  $x = p_{k+1}^2$ , as

$$(13) \quad |\pi(p_{k+1}^2) - \text{li}(p_{k+1}^2)| < \sqrt{\sum_{j=1}^k \frac{l_j \log(p_{j+1}^2/l_j)}{(\log p_{j+1}^2)^2}}.$$

This statement follows from Montgomery and Soundararajan's conjecture discussed in Section 6; the expression on the right side of the inequality is an estimate of the standard deviation of the uncorrelated random model. Due to the negative

contribution to the variance from the covariance terms between intervals, this bound holds stance for the correlated random model and hence the primes. Furthermore, since Montgomery and Soundararajan's conjecture holds empirically, we should expect that the bound should align with the sum of squares of the error functions in each interval,

$$(14) \quad \sum_{j=1}^k (\pi_j - \text{li}_j)^2.$$

We can also express (13) in continuous form. Observe that under the assumption of Cramer's conjecture [Cramér 1936]—that is,  $g_j := p_{j+1} - p_j = O((\log p_j)^2)$ —we have that

$$\log(p_{j+1}^2/l_i) = \log p_{j+1}^2 - \log(2g_j p_{j+1} - g_j^2) \sim \frac{1}{2} \log p_{j+1}^2,$$

which further implies

$$\sum_{j=1}^k \frac{l_j \log(p_{j+1}^2/l_j)}{(\log p_{j+1}^2)^2} \sim \frac{1}{2} \sum_{j=1}^k \frac{l_j}{\log p_{j+1}^2} \sim \frac{1}{2} \text{li}(p_{k+1}^2).$$

We can therefore restate (13) as

$$(15) \quad |\pi(x) - \text{li}(x)| < \sqrt{\frac{1}{2} \text{li}(x)}.$$

Neglecting any constant coefficients, we state this conjecture more generally as

**Conjecture 4.** *The error term in the prime number theorem satisfies*

$$|\pi(x) - \text{li}(x)| = O\left(\sqrt{\text{li}(x)}\right).$$

We continue by plotting the error function  $|\pi(x) - \text{li}(x)|$  together with the different upper bounds, as shown in Figure 12. Note that because Koch's criteria is much larger than the other bounds, we use a logarithmic scale on the  $y$ -axis. As expected, we find that the discrete estimate in (13), corresponding to the uncorrelated random model, lies on top of the sum of squares (14) (in the plot we include the lower order term for higher accuracy), while the error function  $|\pi(x) - \text{li}(x)|$  is placed well below these bounds. The continuous estimate (15) lies close to the discrete estimate (13), but a bit higher, since the lower order terms are not accounted for here; nonetheless, as  $x \rightarrow \infty$ , we expect their relative distance to decrease. Slightly above these curves we find the upper bound of the uncorrelated random model (12). And finally, we observe that Koch's criteria for the Riemann hypothesis is residing high above all the other curves, leaving plenty of room below it.

Essentially then, by considering the prime counting function  $\pi(x)$  as a sum over correlated random variables, we have arrived at a fundamental explanation of why the Riemann hypothesis must be correct, as illustrated by the different theoretical bounds displayed in Figure 12. A purely technical proof still awaits, but perhaps what we have presented here can serve as a stepping stone towards one.

#### REFERENCES

- C Bays and R H Hudson. A new bound for the smallest  $x$  with  $\pi(x) > \text{li}(x)$ . *Math. Comput.*, 69:1285–1296, 2000.

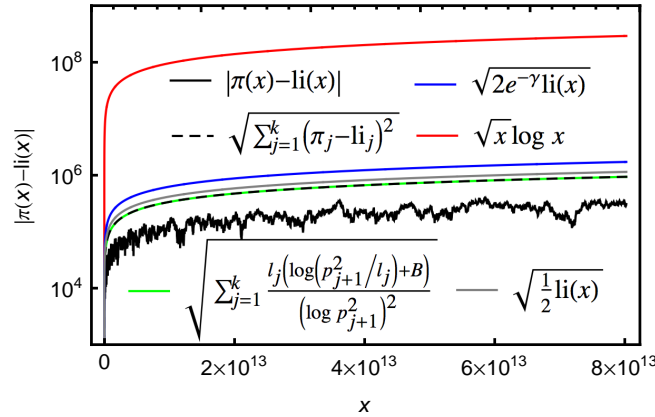


FIGURE 12. The error function  $|\pi(x) - \text{li}(x)|$  and different upper bounds plotted at the values  $x = p_{k+1}^2$  for every 100th value of  $k$ ,  $1 \leq k \leq 6 \times 10^5$ .

- H Cramér. On the Order of Magnitude of the Difference Between Consecutive Prime Numbers. *Acta Arith.*, 2:23–46, 1936.
- C J de la Vallée-Poussin. Recherches analytiques sur la thorie des nombres premiers. *Ann. Soc. Sci. Bruxelles*, 20:183–256, 1899.
- A Granville. Harald Cramér and the distribution of prime numbers. *Scand. Actuarial J.*, 1995(1):12–28, January 1995a.
- A Granville. Unexpected irregularities in the distribution of prime numbers. In *Proceedings of the International Congress of Mathematicians (Zürich, 1994)*, volume 1,2, pages 388–399. Basel, Birkhauser, 1995b.
- M Hausman and H N Shapiro. On the mean square distribution of primitive roots of unity. *Commun. Pure App. Math.*, 26:539–547, 1973.
- D R Heath-Brown. The number of primes in a short interval. *J. Reine Angew. Math.*, 389:22–63, 1988.
- J W Littlewood. Distribution des Nombres Premiers. *C. R. Acad. Sci. Paris*, 158:1869–1872, 1914.
- H Maier. Primes in short intervals. *Michigan Math. J.*, 32(2):221–225, 1985.
- F Mertens. Ein Beitrag zur analytischen Zahlentheorie. *J. Reine Angew. Math.*, 78:46–62, 1874.
- H L Montgomery and K Soundararajan. Primes in short intervals. *Commun. Math. Phys.*, 252:589–617, 2004.
- M Rubinstein and P Sarnak. Chebyshev’s bias. *Exper. Math.*, 3(3):173–197, 1994.
- K Soundararajan. The distribution of prime numbers. In *Equidistribution in number theory, an introduction*, pages 59–83. Springer, Dordrecht, 2007.
- H von Koch. Sur la distribution des nombres premiers. *Acta Math.*, 24(1):159–182, 1901.
- A Zaccagnini. Primes in almost all short intervals. *Acta Arith.*, 84(3):225–244, 1998.

E-mail address: kolbjorn@chalmers.se

COMPLEX SYSTEMS GROUP, DEPARTMENT OF ENERGY AND ENVIRONMENT, CHALMERS UNIVERSITY OF TECHNOLOGY, 41296 GOTHENBURG, SWEDEN